User Analysis Computing at CDF

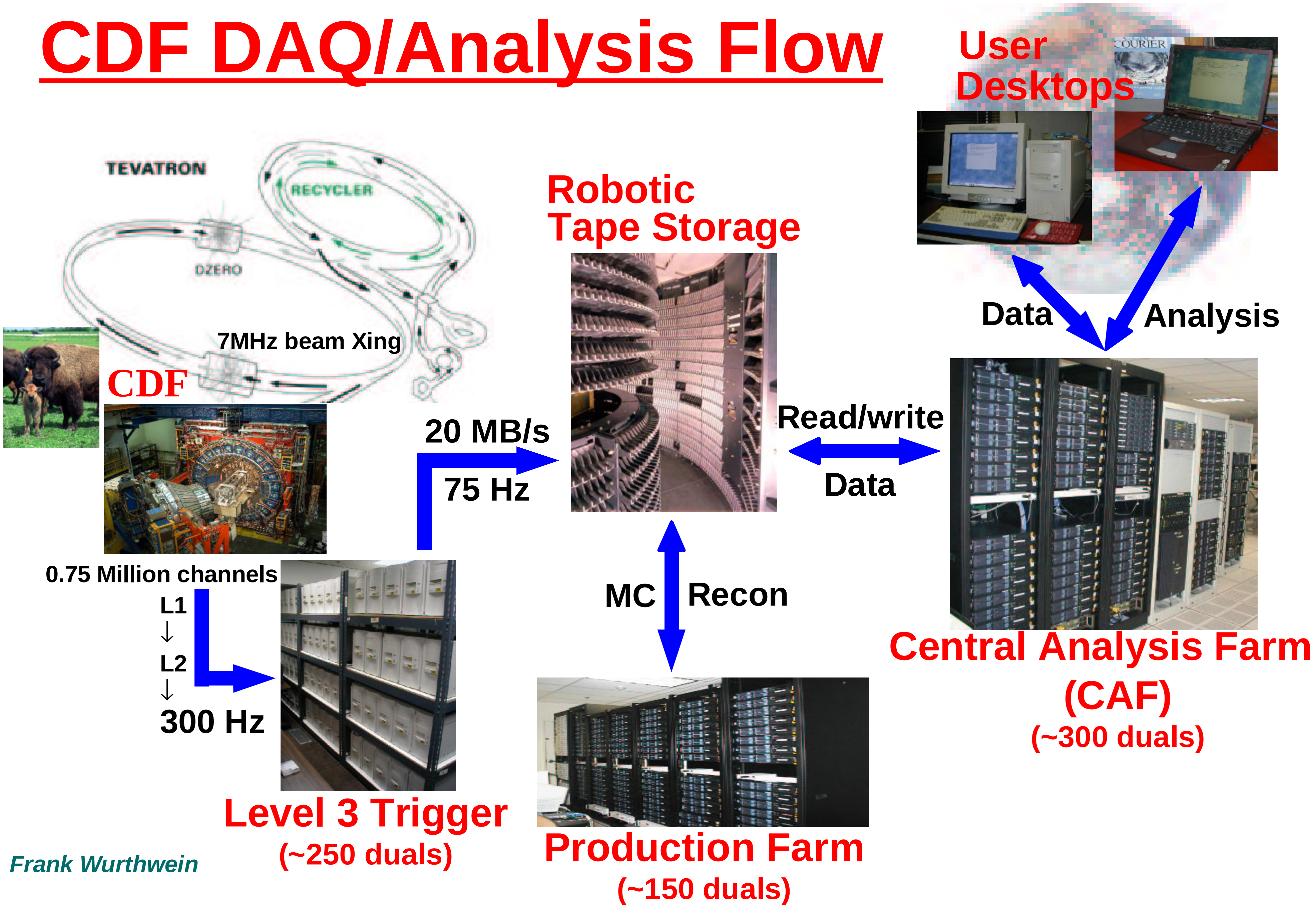
Frank Wurthwein

MIT/UCSD/FNAL-CD

for the CDF Collaboration

- Computing Model
  - status
  - Future directions

CDF DAQ/Analysis Flow



7MHz beam Xing

CDF

0.75 Million channels

L1 ↓ L2 ↓ 300 Hz

Level 3 Trigger

(~250 duals)

20 MB/s

75 Hz

Robotic Tape Storage

MC

Recon

Production Farm

(~150 duals)

Read/write

Data

Central Analysis Farm (CAF)

(~300 duals)

Data

Analysis

User Desktops

Frank Wurthwein

Data/Software Characteristics

Data Characteristics:

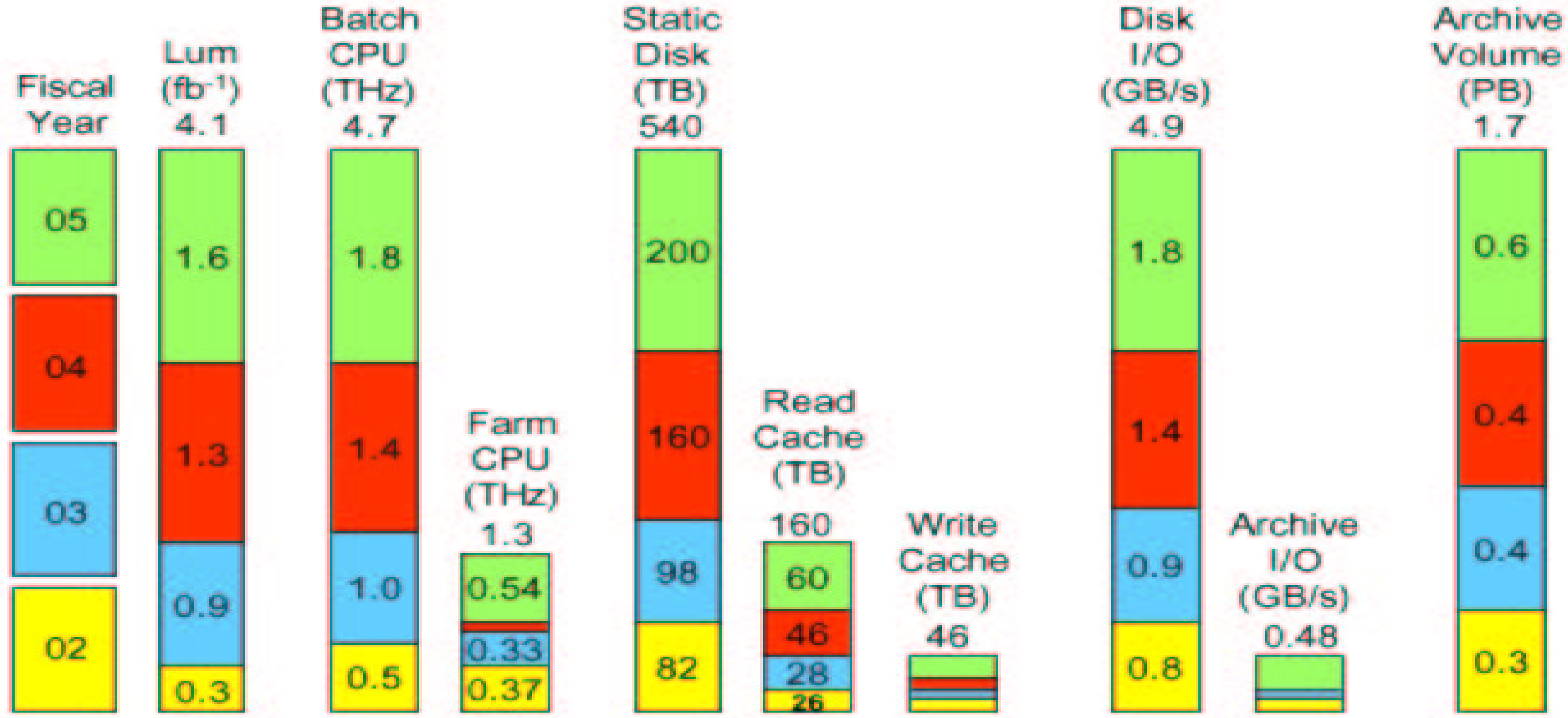
- Root I/O: ~80-400 kB/event (configurable content)
- 'Standard' ntuple: 5-10 kB/event
- Typical RunIIa secondary dataset size: 10^7 events
- Winter03 physics: ~100 datasets adding up to ~50TB
- Largest dataset for Winter03 physics: 3.5e7 evts

Analysis Software:

- Typical analysis jobs run @ 5 Hz on 1 GHz P3

  → few MB/sec
- CPU rather than I/O bound (FastEthernet)

Computing Requirements



Requirements set by goal:

200 simultaneous users to analyze secondary data set (10^7 evts) in a day

Need ~700 TB of disk and ~5 THz of CPU by end of FY'05:

→ need lots of disk → need cheap disk → IDE Raid

→ need lots of CPU → commodity CPU → dual Intel/AMD

Computing Model

Interactive Computing on desktop:

- Complete access to all data from desktop via dCache & rootd

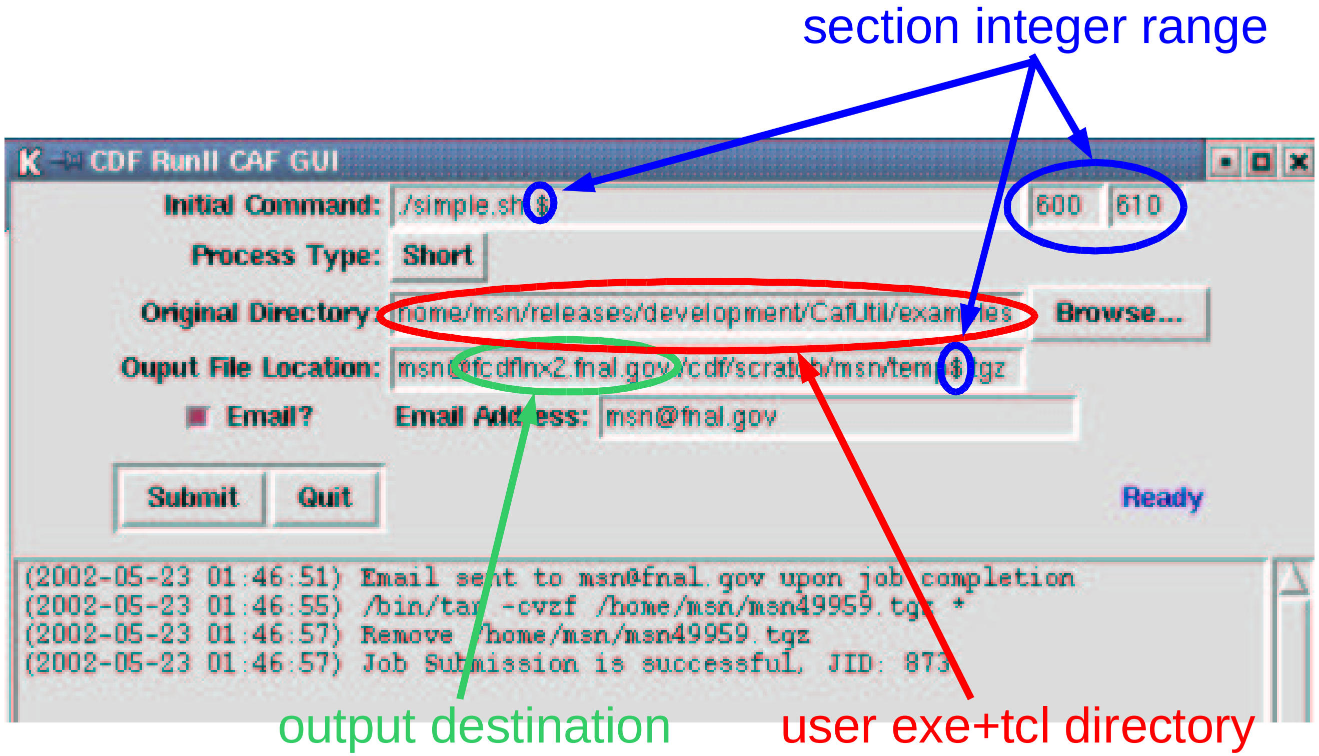
Batch Computing on "remote" cluster(s):

- Binary compatible with desktop
- qsub, qstat, kill, ls, tail, top via command line/web
- Large scale parallelisation with single submission
  - Single summary email upon completion
- User scratch space inside cluster
  - Krb5 ticket created @ launch time
- Data access Winter03: 90% NFS+rootd, 10% dCache

Example job submission

- Compile, build, debug analysis job on 'desktop'

- Fill in appropriate fields & submit job



section integer range

Initial Command: ./simple.sh $ 600 610

Process Type: Short

Original Directory: /home/msn/releases/development/CafUtil/examples

Ouput File Location: msn@fcdflnx2.fnal.gov:/cdf/scratch/msn/temp$.tgz

■ Email? Email Address: msn@fnal.gov

```
(2002-05-23 01:46:51) Email sent to msn@fnal.gov upon job completion
(2002-05-23 01:46:55) /bin/tar -cvzf /home/msn/msn49959.tgz *
(2002-05-23 01:46:57) Remove /home/msn/msn49959.tgz
(2002-05-23 01:46:57) Job Submission is successful, JID: 873
```

output destination

user exe+tcl directory

- Retrieve output using kerberized FTP tools

  ... or write output directly to 'desktop'!

Web Monitoring of User Queues

Each user a different **queue**

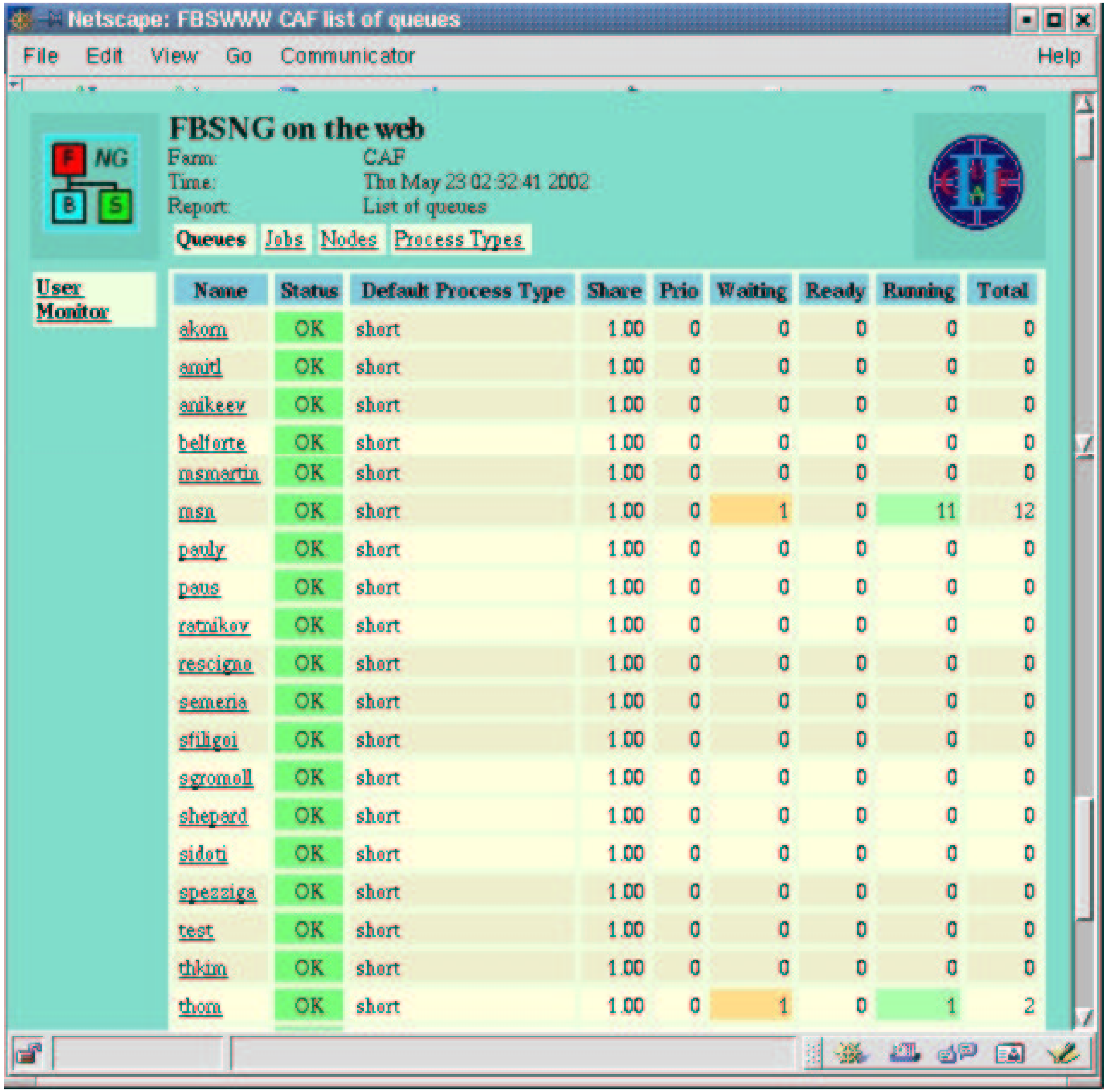
Process type for job length

- **test**: 5 mins
- **short**: 2 hrs
- **medium**: 6 hrs
- **long**: 2 days

This example:

1 job → 11 sections

(+ 1 additional section automatic for job cleanup)



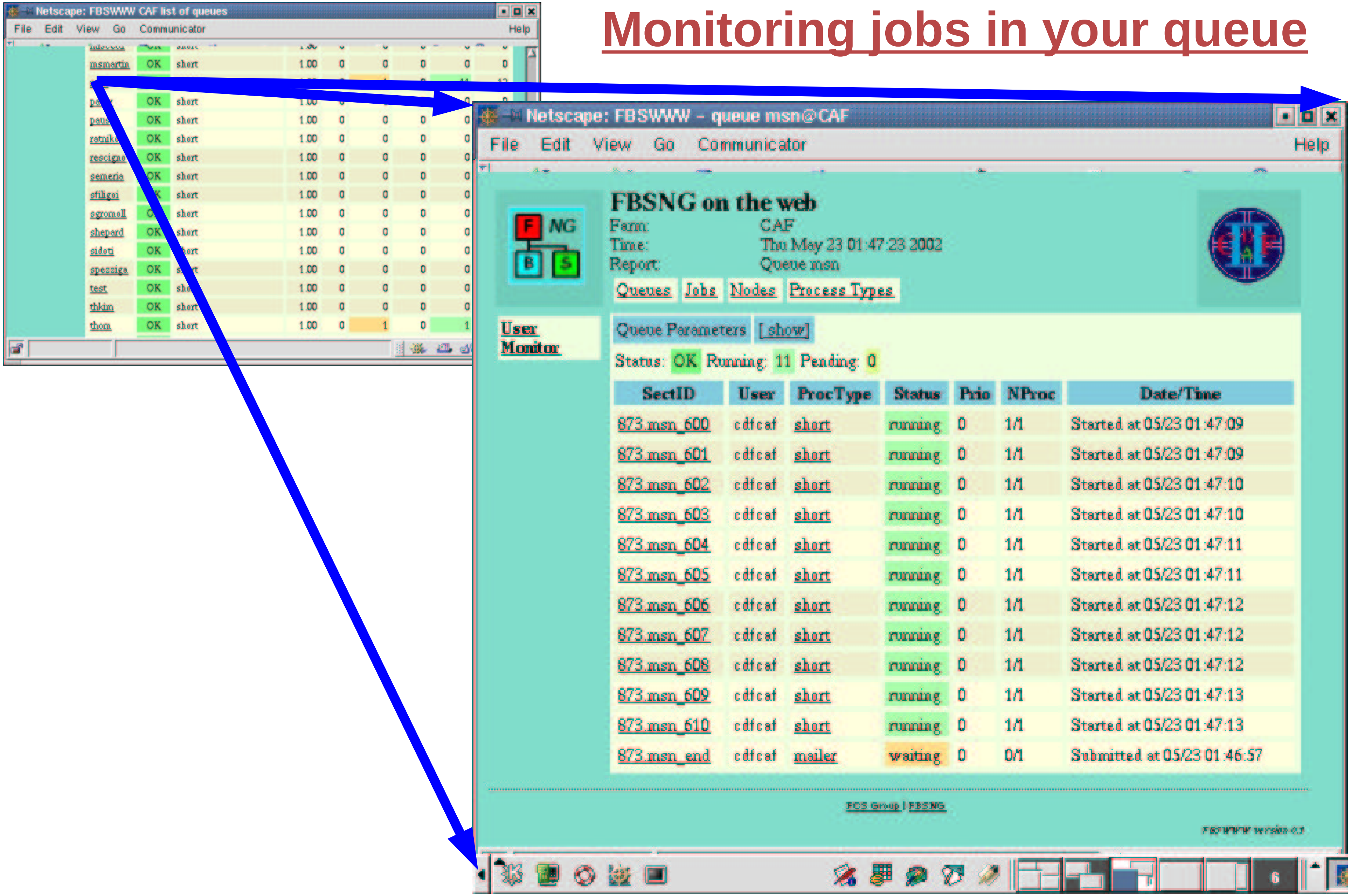
FBSNG on the web

Farm: CAF
Time: Thu May 23 02:32:41 2002
Report: List of queues

Queues | Jobs | Nodes | Process Types

Name	Status	Default Process Type	Share	Prio	Waiting	Ready	Running	Total
akorn	OK	short	1.00	0	0	0	0	0
amitl	OK	short	1.00	0	0	0	0	0
anikeev	OK	short	1.00	0	0	0	0	0
belforte	OK	short	1.00	0	0	0	0	0
msmartin	OK	short	1.00	0	0	0	0	0
msn	OK	short	1.00	0	1	0	11	12
pauly	OK	short	1.00	0	0	0	0	0
paus	OK	short	1.00	0	0	0	0	0
ratnikov	OK	short	1.00	0	0	0	0	0
rescigno	OK	short	1.00	0	0	0	0	0
semeria	OK	short	1.00	0	0	0	0	0
sfiligoi	OK	short	1.00	0	0	0	0	0
sgromoll	OK	short	1.00	0	0	0	0	0
shepard	OK	short	1.00	0	0	0	0	0
sidoti	OK	short	1.00	0	0	0	0	0
spezziga	OK	short	1.00	0	0	0	0	0
test	OK	short	1.00	0	0	0	0	0
thkim	OK	short	1.00	0	0	0	0	0
thom	OK	short	1.00	0	1	0	1	2

Monitoring jobs in your queue



Netscape: FBSWWW CAF list of queues
File Edit View Go Communicator Help

msmartin	OK	short	1.00	0	0	0	0	0
pa	OK	short	1.00				0	0
paus	OK	short	1.00	0	0	0	0	
ratnik	OK	short	1.00	0	0	0	0	
rescigno	OK	short	1.00	0	0	0	0	
semeria	OK	short	1.00	0	0	0	0	
sfiligoi		short	1.00	0	0	0	0	
sgromoll		short	1.00	0	0	0	0	
shepard	OK	short	1.00	0	0	0	0	
sidoti	OK		1.00	0	0	0	0	
spezziga	OK		1.00	0	0	0	0	
test	OK		1.00	0	0	0	0	
thkim	OK	short	1.00	0	0	0	0	
thom	OK	short	1.00	0	1	0	1	

Netscape: FBSWWW - queue msn@CAF
File Edit View Go Communicator Help

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:47:23 2002
Report: Queue msn

Queues | Jobs | Nodes | Process Types

User Monitor

Queue Parameters [show]

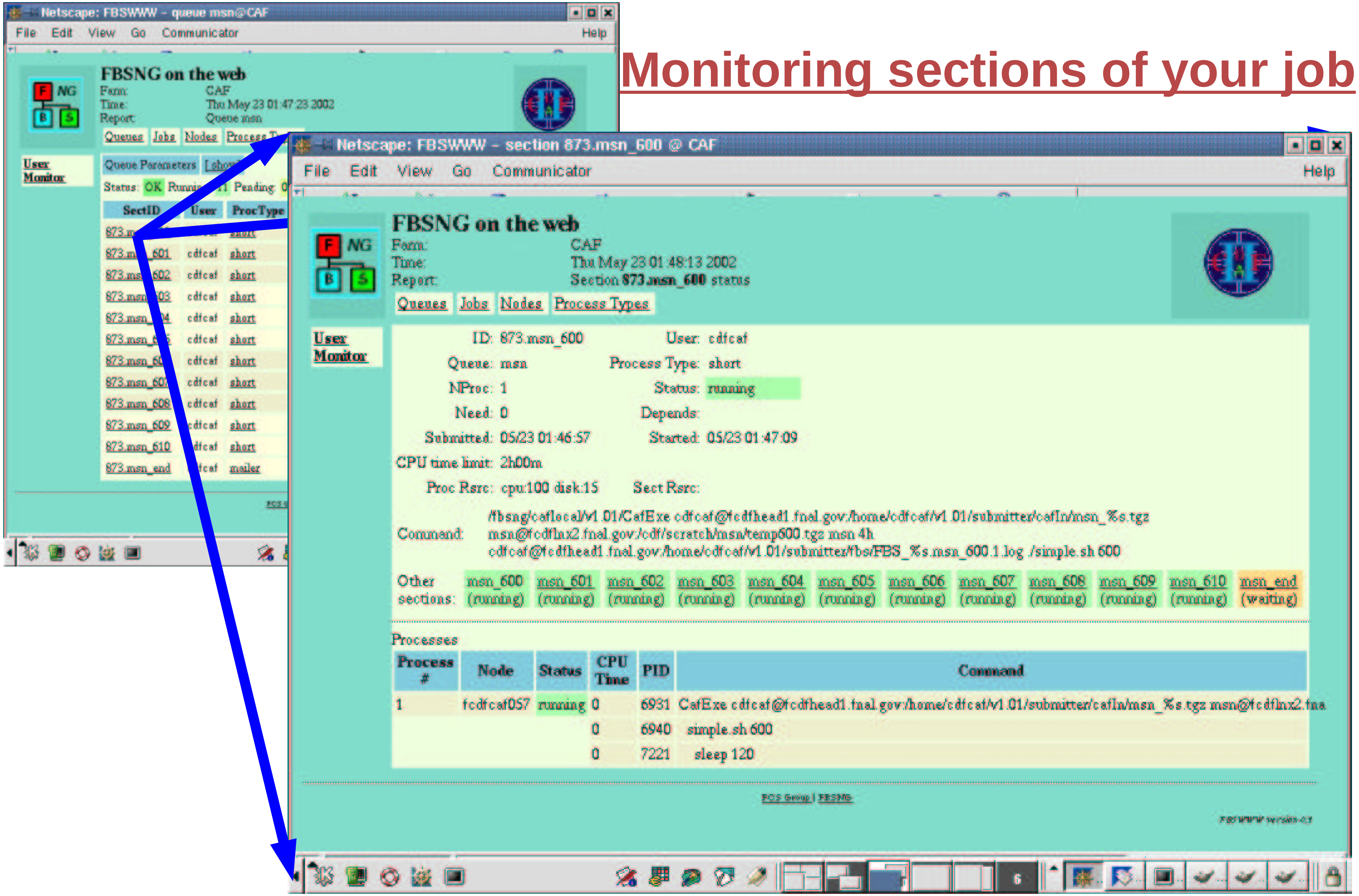
Status: OK Running: 11 Pending: 0

SectID	User	ProcType	Status	Prio	NProc	Date/Time
873.msn_600	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:09
873.msn_601	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:09
873.msn_602	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:10
873.msn_603	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:10
873.msn_604	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:11
873.msn_605	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:11
873.msn_606	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_607	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_608	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:12
873.msn_609	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:13
873.msn_610	cdfcaf	short	running	0	1/1	Started at 05/23 01:47:13
873.msn_end	cdfcaf	mailer	waiting	0	0/1	Submitted at 05/23 01:46:57

FCS Group | FBSNG

FBSWWW version [illegible]

Monitoring sections of your job



Netscape: FBSWWW - queue msn@CAF

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:47:23 2002
Report: Queue msn

Queues Jobs Nodes Process T

User Monitor

Queue Parameters [show

Status: OK Runnin 11 Pending: 0

SectID	User	ProcType
873.m		short
873.m 601	cdfcaf	short
873.ms 602	cdfcaf	short
873.msn 03	cdfcaf	short
873.msn 4	cdfcaf	short
873.msn 5	cdfcaf	short
873.msn 60	cdfcaf	short
873.msn 607	cdfcaf	short
873.msn_608	cdfcaf	short
873.msn_609	cdfcaf	short
873.msn_610	dfcaf	short
873.msn_end	fcaf	mailer

Netscape: FBSWWW - section 873.msn_600 @ CAF

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:48:13 2002
Report: Section 873.msn_600 status

Queues Jobs Nodes Process Types

User Monitor

ID:	873.msn_600	User:	cdfcaf
Queue:	msn	Process Type:	short
NProc:	1	Status:	running
Need:	0	Depends:	
Submitted:	05/23 01:46:57	Started:	05/23 01:47:09
CPU time limit:	2h00m		
Proc Rsrc:	cpu:100 disk:15	Sect Rsrc:	

Command:
```
/fbsng/caflocal/v1.01/CafExe cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/cafIn/msn_%s.tgz
msn@fcdflnx2.fnal.gov:/cdf/scratch/msn/temp600.tgz msn 4h
cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/fbs/FBS_%s.msn_600.1.log ./simple.sh 600
```

Other sections: msn_600 (running) msn_601 (running) msn_602 (running) msn_603 (running) msn_604 (running) msn_605 (running) msn_606 (running) msn_607 (running) msn_608 (running) msn_609 (running) msn_610 (running) msn_end (waiting)

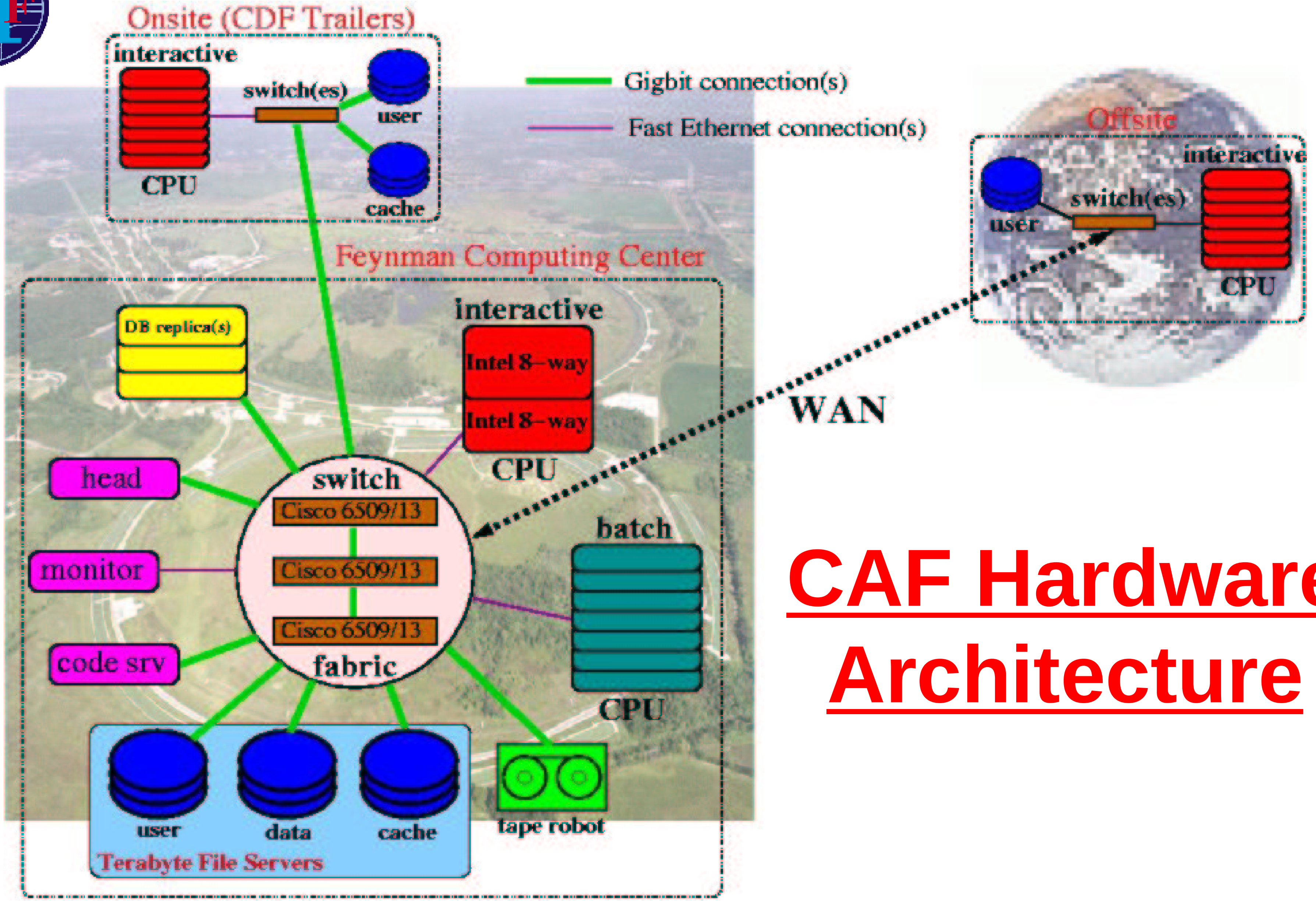
Processes

Process #	Node	Status	CPU Time	PID	Command
1	fcdfcaf057	running	0	6931	CafExe cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/cafIn/msn_%s.tgz msn@fcdflnx2.fna
			0	6940	simple.sh 600
			0	7221	sleep 120

FCS Group | FBSNG

FBSWWW version-0.1

CAF Hardware Architecture



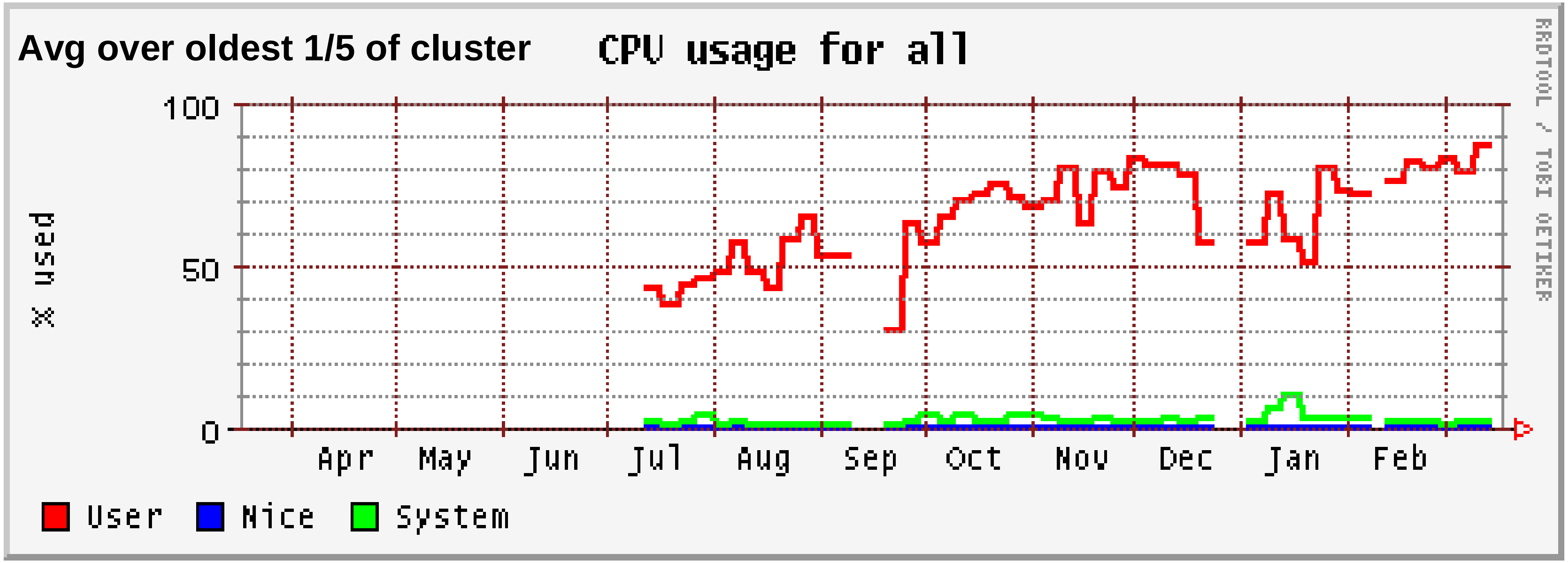
CAF utilization

User perspective:

- 10,000 jobs launched/day
- 400 users total
- 100 users per day

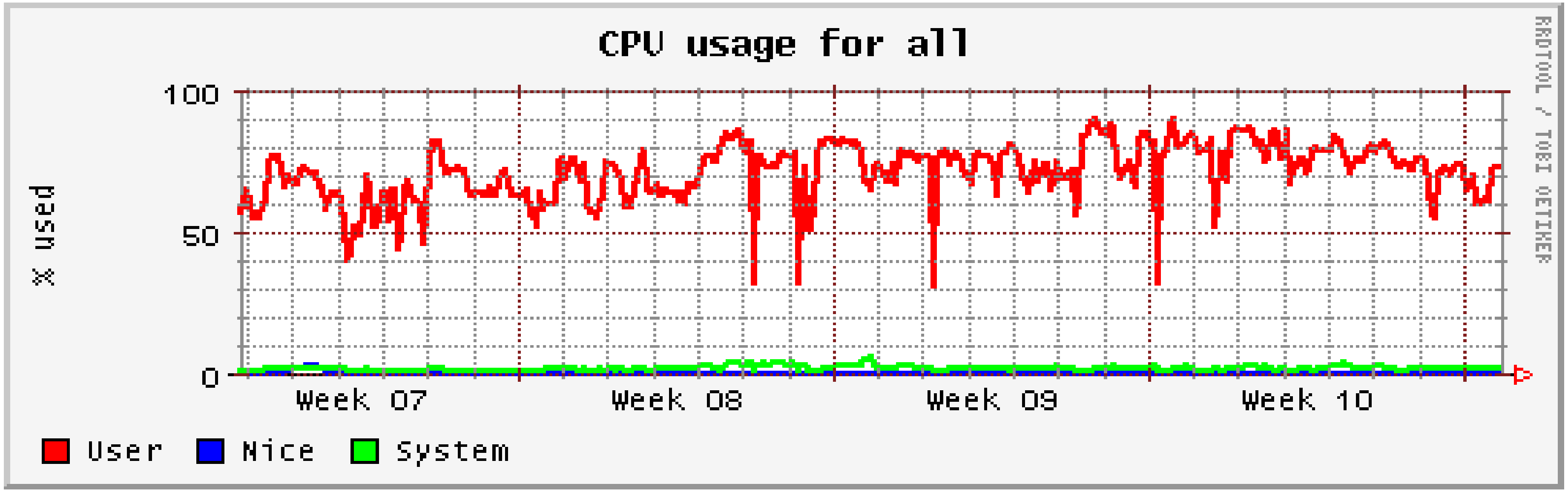
System perspective:

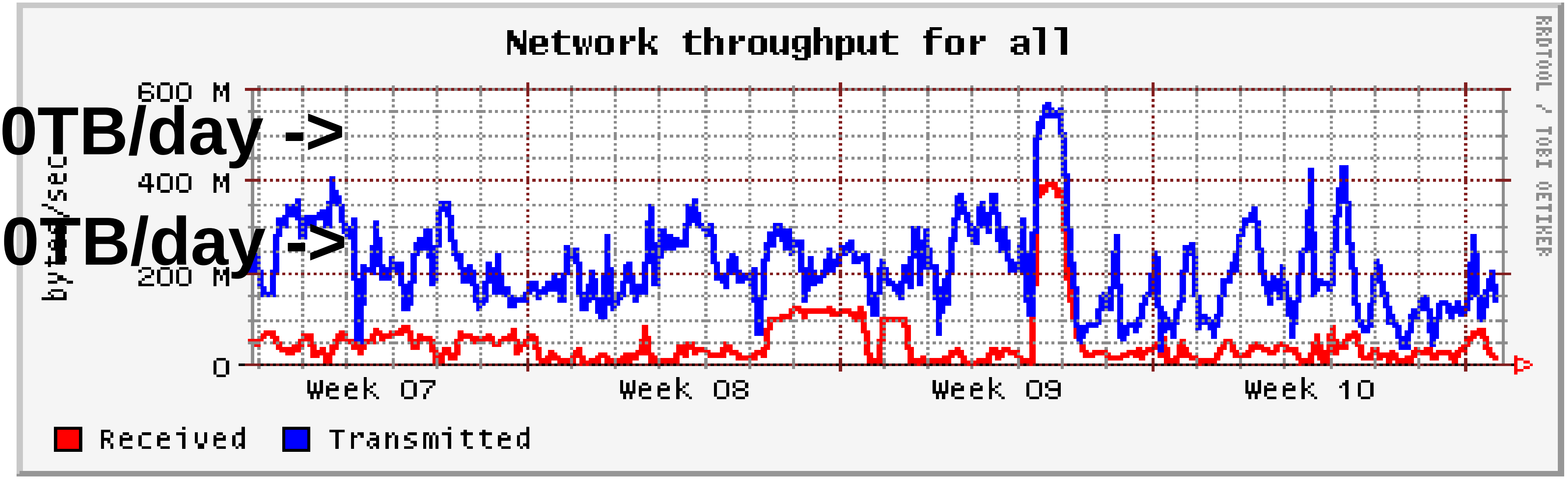
- Up to 90% avg CPU utilization
- 200-600MB/sec I/O
- Failure rate ~1/2000
- Avg uptime of WN = 60days



Avg over oldest 1/5 of cluster

CAF utilization last month





40TB/day ->

20TB/day ->

Status @ FNAL Today

User analysis computing based on commodity PC's

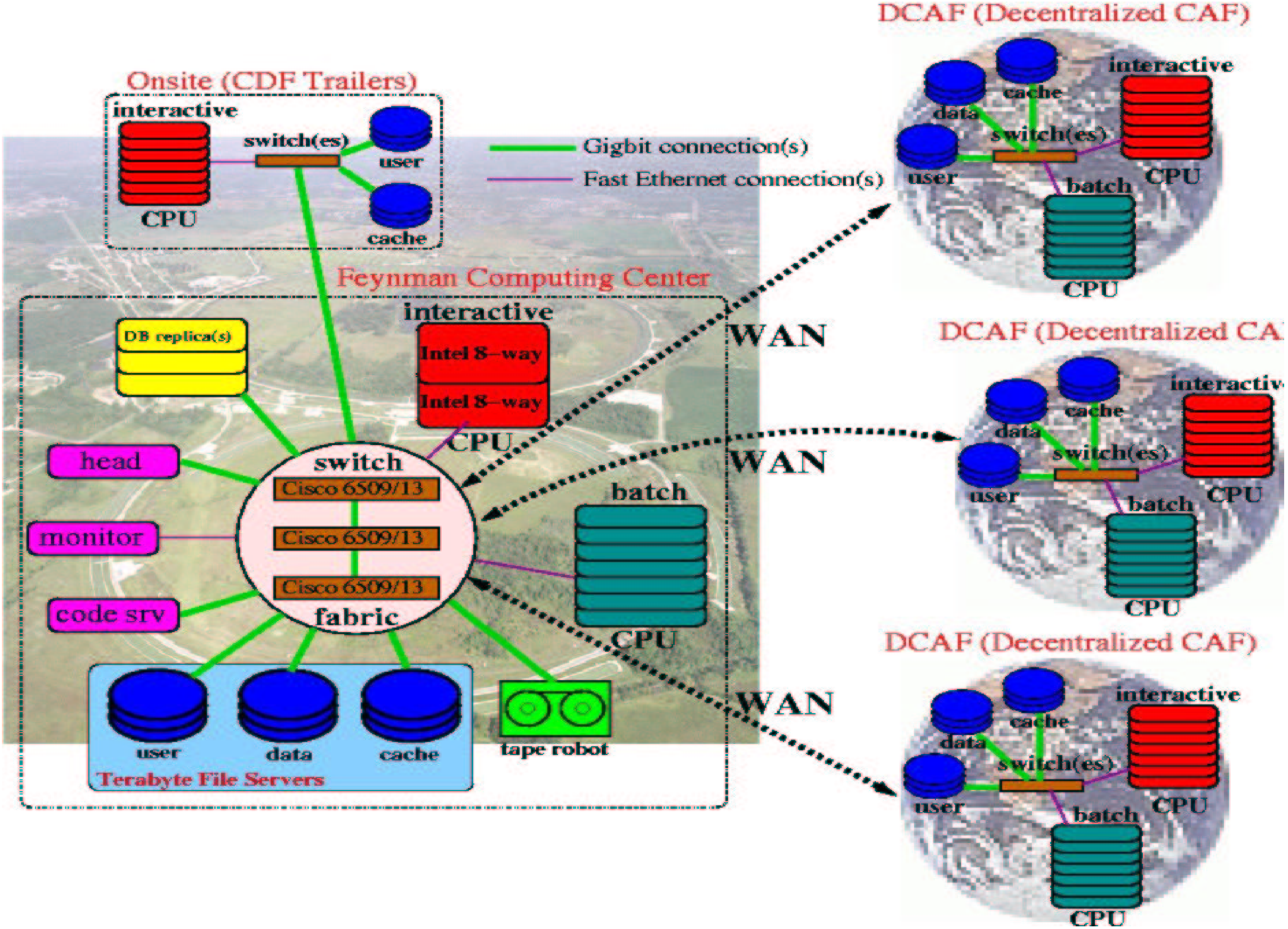
180TB disk space   1THz batch CPU

Focus on building strong infrastructure

up to 600MB/sec I/O   99.95% reliability

that has been deployed as part of CDF grid
"proof of principle" for SC2002 demo.

Future Directions



CDF grid = 3 pieces

CAF:

- Local cluster management
- Remote submission
- Fully in production: 99.95% reliability

SAM:

- WAN capable DH system
- Use for remote MC production in summer 03

JIM:

- Grid broker (based on condor,globus,sam)
- Proof of principle fall 02

Summary & Conclusions

CDF's computing model makes offsite computing contributions possible.

Offsite contributions, and accounting thereof is desirable.

The devil is in the detail.